EARLY FUSION OF SPARSE CLASSIFICATION AND GMM FOR NOISE ROBUST ASR

Yang Sun, Jort F. Gemmeke, Bert Cranen, Louis ten Bosch, Lou Boves

Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
{Y.Sun, J.Gemmeke, B.Cranen, L.tenBosch, L.Boves}@let.ru.nl

ABSTRACT

In previous work we have shown that an ASR system consisting of a dual-input DBN which simultaneously observes MFCC acoustic features and predicted phone labels that are generated by an exemplar-based Sparse Classification (SC) system can achieve better word recognition accuracies in noise than a system observing only one of those input streams. This paper explores two modifications of the SC input to further improve the noise robustness of the dual-input DBN system: 1) integrating more time context and 2) using N best states. Experiments on AURORA-2 reveal that the first approach significantly improves the recognition results at almost all SNRs, but particularly in the more noisy conditions, achieving up to 6.1% (absolute) accuracy gain at SNR -5 dB. The second modification shows that there is an optimal N which allows the maximum attainable accuracy to be even further improved with another 11.8% at -5 dB.

1. INTRODUCTION

Systems based on Hidden Markov models (HMMs), that obtain state likelihoods by modeling feature observations with Gaussian Mixture Models (GMMs), have dominated the automatic speech recognition (ASR) field for the last 30 years. While quite successful in dealing with clean, read or prepared speech, the performance of this type of recognizer is known to degrade dramatically under noisy conditions or with spontaneous conversational speech. Numerous adaptations to the speech pre-processing, the acoustic models, or the decoding algorithms have been proposed to make ASR more robust against variation that was not present in the training data [1]. However, no approach has yet appeared sufficiently powerful to approximate the performance levels of Human Speech Recognition (HSR) (e.g. [2, 3]). There is a growing consensus that besides the stream of acoustic input features (usually MFCC coefficients) which are modeled by GMMs, complementary information is needed. Many studies, such as [4], have shown that hybrid systems which combine a conventional GMM-based system with an additional system that processes the speech data differently, often perform better than either individual system alone and therefore might eventually be able to bridge the gap between ASR and HSR. While features, probability estimates, or system outputs can be combined at many levels (e.g. [5]), our work aims at very early fusion, such that we can eventually exploit low-level dependencies between features.

In previous work [6], we demonstrated that a dual input dynamic Bayesian network (DBN) which combines the likelihoods from MFCC-based GMMs and predictions of the most likely phone from an exemplar based classifier [dubbed Sparse Classification (SC)], yields good performance over a wide range of SNRs. In this paper, we propose two modifications of the SC input stream to investigate whether we can further improve the performance of the DBN system. In [6] the SC-input was obtained using exemplars that span 10 frames. In [7] it was shown, however, that the recognition performance of an SC decoder can be significantly improved at low SNRs by using larger exemplar sizes, be it at the cost of lower accuracies at high SNRs. In this paper we first investigate whether the dual-input DBN system also benefits from using larger exemplar sizes in low SNRs, without the performance in clean conditions being affected. Second, we use a vector of the N best posterior probability estimates generated by the SC system as the second input stream instead of the label of the most likely state. Adding more information about the likelihood of other states aims at increasing the search space during decoding in such a way that alternative, potentially winning hypotheses (according to the SC system) are enabled that otherwise would fall out. As in [6], we evaluate the effectiveness of these modifications using the AURORA-2 task.

The rest of this paper is organized as follows. In Section 2, the dual-input dynamic Bayesian network (DBN) architecture and its two input streams are described. The experimental setup and is described in Section 3, while the two experiments with the two proposed modifications are described and discussed in Sections 4 and 5, respectively. Finally, we present our conclusions in Section 6.

2. DUAL-INPUT DBN

2.1 DBN architecture

Figure 1 depicts the input stage of the dual-input DBN architecture used in our study. The discrete, random variable s_t represents the states over time t and the shaded circular nodes x_t represent the traditional, continuous MFCC features modeled by GMMs (*GMM* hereafter). The square (hidden) nodes SC_t represent some additional external evidence, in our case provided by the SC system (cf. Section 2.2).

In this study GMM and the SC are assumed to be conditionally independent, i.e., we approximate the joint probability $p(s_t, x_t, SC_t) \approx p(x_t|s_t) \cdot p(SC_t|s_t) \cdot p(s_t)$. With D representing the total number of states, the first term is modeled by D Gaussian mixture models GMMs, while the second term is a conditional probability table (CPT) of size $D \times D$. Both $p(x_t|s_t)$ and $p(SC_t|s_t)$ are trained on clean speech while simultaneously presenting the MFCC and SC inputs to the DBN.

Instead of being directly observed as in [6], the values of SC_t are determined in an indirect way by means of virtual evidence (VE) nodes VE_t (see also [8]). For each time frame, the virtual evidence is supplied to the DBN in the form of an 1xD CPT (on the arc between VE_t and SC_t). Thus, in practice, this VE vector gives the SC_t variable a prior distribution

The research of Yang Sun has received funding from European Community's Seventh Framework Programme [FP7/2007-2013] under grant agreement no. 213850 - SCALE. The research of Jort F. Gemmeke is supported by the Dutch-Flemish STEVIN Program.

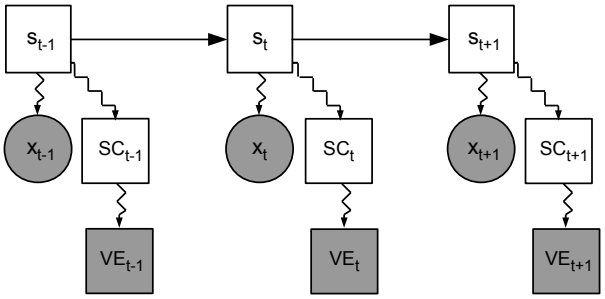


Figure 1: Input stage of the dual input DBN.

which corresponds to the externally computed probability estimates from the SC system for each of the D states.

2.2 Sparse Classification

In the Sparse Classification (SC) system [7], an observed speech spectrogram of length T frames is expressed as a sparse, non-negative linear combination of segments of speech, named *exemplars*. In this work, we compared two exemplar sizes, $T = 10$ and $T = 30$ frames, denoted by *T10* and *T30*, respectively, hereafter. Likewise, segments of noise are modelled as a linear combination of noise exemplars. Using a collection of noise and speech exemplars, called a *dictionary*, each speech segment (clean or noisy) is modelled as a linear combination of both speech and noise exemplars. By finding the sparsest possible set of speech and noise exemplars that approximates the observed noisy speech, we obtain a sparse representation of each observed speech segment.

Each frame of each exemplar in the speech part of the dictionary is labelled with HMM-state labels obtained from a forced alignment with a conventional MFCC/GMM-based decoder. Using the recovered sparse representation, we use the weight of each speech exemplar to obtain estimates of the posterior state likelihoods by calculating the weighted linear combination of underlying state labels.

3. EXPERIMENTAL SETUP

The MFCC inputs to the DBN used in training were obtained from the clean training set of the AURORA-2 corpus (8440 utterances). They consists of 39 dimensional vectors containing 12 MFCC features, plus a separate log-energy coefficient as well as the corresponding first and second order delta coefficients. They are based on a 23 band Mel frequency spectrum using a frame shift of 10ms and a window length of 25ms. Subsequently, the MFCCs are normalised with respect to their mean and variance per utterance. The statistics of the MFCC feature vectors are modelled by mixtures of diagonal covariance Gaussians. Our final GMMs consists of mixtures of up to 32 diagonal covariance Gaussians.

To obtain the SC information, we used the same configuration as in [7]. In a nutshell, the SC method operates on 23-dimensional Mel-scale magnitude features, and uses a dictionary created with 4000 clean exemplars, randomly extracted from the multi-condition training set and 4000 noise exemplars also randomly selected from the multi-condition training set (by subtracting the corresponding clean speech signals). We used two exemplar sizes *T10* or *T30*. The output of the SC system is a 179 dimensional vector for each MFCC frame, corresponding to the posterior probability estimate of each HMM-state, of which the dimensionality is then reduced by only retaining the N most likely elements before it is presented to the DBN. The total probability mass of this vector is then normalized so as to add up to one. In order to keep correspondence with the original SC system in [7], which employs Viterbi decoding directly on the posterior probability estimates of the classifier, we decided not to train the CPT that models the relation between SC_t and s_t, but to replace it with an identity matrix of which the off-diagonal elements were given a floor value of 1e-30.

For the first experiment (of which the details are described in Section 4), we only used test set 'A', which consists of utterances containing a sequence of one to seven connected digits (taken from the set 'zero-nine' and 'oh') which have been artificially corrupted with four noise types (subway, car, babble, exhibition hall) yielding seven noise levels, viz. clean, and SNR = 20, 15, 10, 5, 0, and -5 dB. For the second experiment (described in Section 5), we also used test set 'B', which contains four other noise types (restaurant, street, airport, train station). These noise types are not covered by the noise dictionary employed by SC and thus allows investigation of performance in mismatched conditions.

Instead of just reporting the results of the dual-input system (denoted by *GMM/SC* hereafter), we also provide results of each single input (SC-only and GMM-only) system separately. These data will help us to understand the contributions of the two input streams (and their interactions) in obtaining the recognition results. The word recognition accuracies are averaged over the four noise types (either in set 'A' or 'B') in all results. In experiment 1 (cf. section 4) we also average the accuracies for the SNRs between 0 and 20 dB, while we report the results for clean and SNR = -5 dB separately.

4. EXPERIMENT 1: INCREASING THE AMOUNT OF CONTEXTUAL INFORMATION

In [6] we successfully used a dual-input DBN to achieve a more noise robust speech recognition by combining MFCC-based GMM and *T10* SC input. In [7] it was found, however, that a Viterbi decoder using the state probability estimates obtained with an SC system using exemplar sizes of 30 instead of 10 frames, gives significantly higher word accuracies at low SNRs, be it at the cost of a decreased performance at high SNRs. The goal of this first experiment is to investigate to what extent the increased noise robustness at low SNRs using *T30* SC input can be exploited in a dual-input system without reducing the accuracy at high SNRs, which is mainly due to the GMM input. In order keep results maximally comparable with those obtained in [6], we will first focus on the differences in recognition performance when the GMM scores are combined with the identity of the most likely (state) candidate according to the SC-system.

Results for the AURORA-2 test set A are shown in Table 1. The first row [$GMM(base)$] shows the word accuracies obtained with the GMM-only system. Since the concept of exemplar size is meaningless for this system, the results are simply replicated under $T10$ and $T30$. The second row [$SC(base)$] shows the recognition results for the SC-only baseline system using the full 179-dimensional probability vector from the SC-system when the same Viterbi decoder is used as in [7]. Comparison of the rows confirms earlier observations that (1) the GMM-system is superior in the clean condition, while in the very noisy condition the SC-system outperforms the GMM-system [6] and (2) the $T30$ $SC(base)$ system outperforms the $T10$ system in very noisy conditions at the cost of the performance in clean conditions [7].

When we modify the original SC system by only using the maximum likely candidate, the results in the third

Table 1: *Word accuracy for test set A in %.* GMM(base) *and* SC(base) *refer to the two baseline systems in which the full information from either the GMM or the SC stream is used respectively.* SC(base-1) *refers to the SC baseline system that only uses one dimensional state probability estimates* SC(1) *refers to the DBN system that only uses one dimensional SC input.* GMM/SC(1) *is the dual input system with both GMM and one dimensional SC input.*

	T10			T30		
SNR(dB)	cln	0-20	-5	cln	0-20	-5
GMM(base)	99.4	84.0	26.4	99.4	84.0	26.4
SC(base)	96.4	87.6	**42.2**	93.7	88.5	**57.1**
SC(base-1)	73.5	47.9	-14.3	90.9	81.9	45.2
SC(1)	72.7	47.0	-14.5	90.5	81.4	44.7
GMM/SC(1)	**99.4**	**87.9**	32.8	**99.6**	**89.8**	38.9

row [*SC(base-1)*] are obtained. As expected the overall performance goes down, particularly in the $T10$ case. This phenomenon indicates that the other elements of the SC-probability vector *do* contain information which is important for decoding. Despite the fact that only the maximum likely candidate is used, at SNR=-5 dB $T30$ yields even higher word accuracies than $T10$ using the full sized input vector.

The fourth row in Table 1, marked as $SC(1)$ shows the results for the DBN system using only the 1 dimensional SC-input. The results in this row allow us to verify that our single stream DBN system performs comparably to the Viterbi system that was used to obtain the results in the previous rows. As can be seen the results obtained with both systems are very similar, with small differences that can be attributed to the differences between the back-end of the Viterbi decoder and the DBN system, such as pruning and word-transition probabilities.

Finally in the fifth row, the results are shown when the GMM and the 1 dimensional SC input are combined. Clearly, also for the combined system the performance is better for the $T30$ than for the $T10$ SC input. This holds for all noise conditions, indicating that incorporating more time context is a powerful mechanism to increase noise robustness. Also noteworthy is that both in the clean condition and noisy conditions (except the SNR=-5 dB case), the dual input system outperforms the best of the single input systems $GMM(base)$ and $SC(1)$.

Since in the results in Table 1, it was observed that there is a loss of performance for the original SC-system by reducing the dimensionality to the most likely state, we investigate in the next experiment whether increasing the dimensionality of the SC input can further strengthen the combination.

5. FROM THE BEST TO N-BEST STATE PROBABILITY VECTOR

5.1 SC-only input

In the previous experiments the additional input to the DBN consisted of the single state that was most likely according to the SC system. Shrinking the vector with probability estimates from the SC system from 179 dimensions into a single label implies that we rely completely on the information in the top prediction made by SC system. By flooring the remaining 178 state likelihoods (to 1e-30) we ensured that it is always possible for the Viterbi search to find a path from beginning to end. However, it is obvious that a one dimensional SC input cannot always be correct, especially at low SNRs.

To explore the influence of using higher-dimensional SC input, we first did the following experiment. We investigated a DBN system which is fed solely with a vector of probability estimates from an SC system using $T30$ exemplars. Of this vector only the largest $N \geq 1$ elements are kept after which the vector is normalized to have a total probability mass of one. The word accuracies for this SC-only system for varying N are shown in Figure 2.

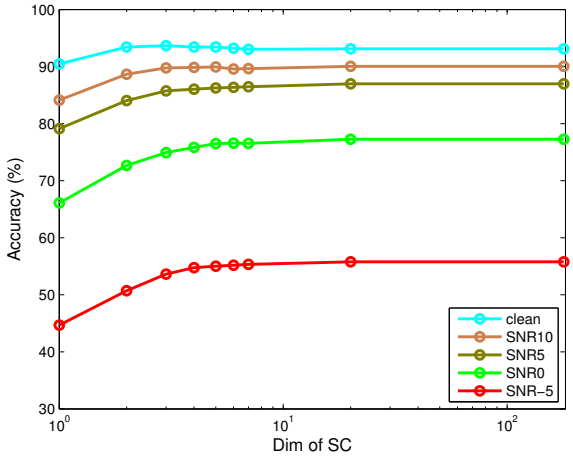
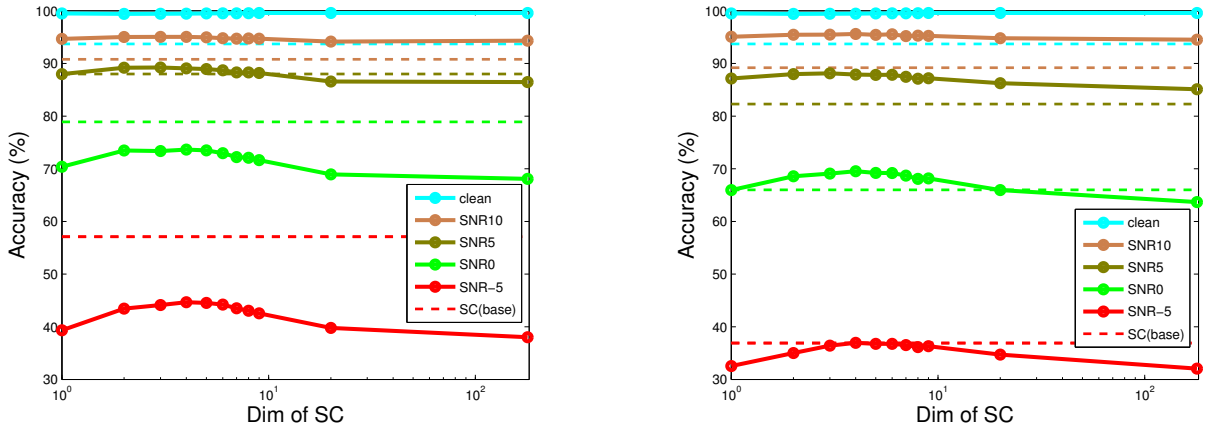


Figure 2: Word accuracy of SC-only $T30$ system at different SNRs on AURORA-2 testing set A as a function of dimension of SC input. The horizontal axis is on a logarithmic scale.

In Figure 2, we can observe that for clean speech, the accuracy goes up from 90.9% at $N = 1$ to 93.4% at $N = 2$. For $N > 2$ the accuracy declines somewhat before converging to 93.1% at $N = 179$. For noisier speech, we observe that the optimal number of SC dimensions is larger. For example, at SNR -5 dB, the accuracy goes up from 44.7% at $N = 1$ to 53.7% at $N = 7$, after which the performance stays constant for larger N.

From this we can conclude that in general, a one dimensional SC input yields suboptimal results and we can gain substantially higher accuracies by using more dimensions of the SC vector. In practice the accuracy often converges for $N > 10$. Using the complete SC vector with probability estimates for all states, however, does not always lead to the best result. The fact that in the clean condition, there exists an optimum for N=2 is due to the fact that adding too many runner-up states may also introduce misleading information. Particularly in the clean condition, the SC-system will yield probability vectors which assign high probabilities to a very limited number of states (i.e. the the SC probability estimates are *sparse*); the rest of the probability estimates will mainly constitute 'noise'. As a consequence, an increase of the number of dimensions N in such condition will initially enable these noisy probability estimates (which will only have very small values) to slightly decrease the recognition performance because it opens up 'wrong' {frame,state}-paths in the search. Beyond a certain value, however, (judging from Figure 2 for $N \geq 10$), the accuracy stays more or less constant because in that range adding more runner-up states does not add any information at all. For more noisy conditions, it can be seen that a larger N is needed before the optimal recognition accuracy is reached. This makes sense if one realizes that the probability estimates in the SC vector at lower SNRs are distributed over more states (i.e. are less sparse). In these condi-



(a) test set A

(b) test set B

Figure 3: Word accuracy of the dual-input DBN on AURORA-2 as a function of state-likelihood-dimensionality of SC input. The left hand panel (a) shows the data for test set A, the right hand panel for test set B. Note that the horizontal axis is displayed logarithmically. The data point on the far right represents 179 likelihood dimensions.

tions, an increase of N will add the probability of runner-up states that do contain real information. For even larger N the same saturation effect is expected as in the clean condition.

5.2 Dual-input system

In a subsequent experiment, we explored the influence of the dimensionality of the SC input in the dual-input DBN. To investigate the impact of noise types not included in the noise dictionary employed by SC, we tested on both test set A and B of AURORA-2. The results are shown in Figure 3a and 3b, respectively.

Comparing the results in Figure 3a for test set A with the results obtained above for the SC-only system (cf. Figure 2) it can be seen that the increase in dimensionality does not have the same effect as when combining the GMM and SC streams. For clean speech, the best accuracy is 99.6%; the differences in accuracy for different N are not statistically significant. At lower SNRs, there is an optimum at $N = 4$, and adding more dimensions has a detrimental effect. For example, at SNR -5 dB, the accuracy first increases from 39.3% at $N = 1$ to 44.7% at $N = 4$, and eventually drops to 39.8% at $N = 179$.

The clear optimum in the dimensionality of the SC input at lower SNRs, not observed when using only a single SC input stream, must come from the interaction of the information between the SC input stream and the GMM input stream. This is most likely due to the fact that adding multiple SC-input dimensions opens up the search space: More states getting a non-zero probability allow more GMM information to be used. However, as the GMM probability estimation of a system trained on clean speech is typically very uninformative at such low SNRs, the use of a larger search space only adds 'noise' to the probabilities. Thus, in contrast to the SC-only system, there will not be a saturation effect as in Figure 2, but rather a decrease in performance due to the misinformation provided by the GMMs.

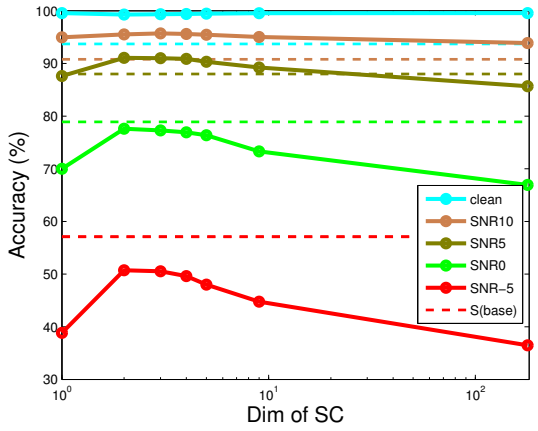
When comparing the results on test set A and B (cf. Figures 3a and 3b, respectively), the main difference is that the absolute performance on test set B is lower than on test set A. Obviously, this is due to the fact that the SC system has a noise dictionary that contains noise types that no longer match the test set, as discussed in [7]. In all other aspects the curves in both figures are very similar: For clean speech, there is hardly any influence of using dimensions $N > 1$, while at lower SNRs, there is a clear optimum at $N = 4$, with the $N = 179$ performance comparable or somewhat lower than the $N = 1$ performance.

Finally, comparing the performance of the dual-input DBN with that of the original, *SC*(*base*) Viterbi-based SC system (indicated by dotted lines of the same colour in Figures 3a and 3b), we can observe that on test set A, the combined system performs much better at SNRs > 0 dB, only losing performance in the noisiest conditions. Compared to a MFCC/GMM-based recogniser, *GMM(base)* in Table 1, we increase the noise robustness substantially without losing any performance on the clean speech. On test set B, the dual-input DBN does as good or better than the original SC system in noisy conditions, without losing any of the clean speech performance.

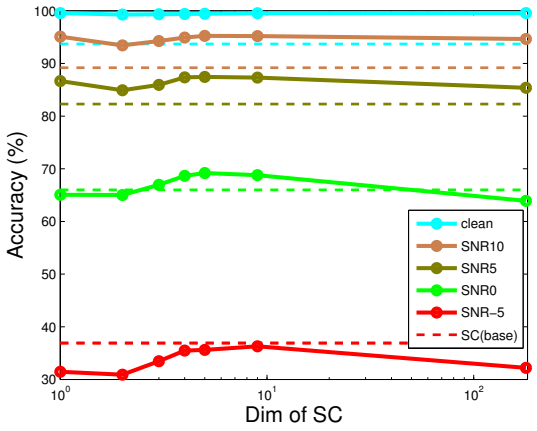
5.3 Trained CPTs

In a final experiment, we *did* train the CPT which accounts for the distribution of the SC probability estimates during training on clean speech, in contrast to the previous experiments which employed a fixed, diagonal CPT. In this experiment, the CPT was trained jointly with the GMM and decoder parameter training, and a separate CPT was trained for every state-dimension N. Training the CPT used for SC input allows the DBN to account for irregularities in the SC probability estimates. The results are shown in Figures 4a and 4b.

From the results we can observe that the recognition accuracies at lower SNRs generally improve: for example on SNR -5 dB on test set A, the dual-input DBN now achieves a recognition accuracy of 50.7 % at $N = 2$. This means that the trained CPT, although trained on clean speech, manages to compensate for (ir)regularities in the SC input. This can be understood as follows. If in the linear combination of exemplars, some exemplars are consequently activated in a certain state, but the corresponding state label of the exemplars are of a different state, the SC system produces an incorrect probability estimate from which a DBN with diagonal CPT cannot recover. The use of a trained CPT allows for to indirectly modelling the relation between exemplar activations



(a) test set A



(b) test set B

Figure 4: Word accuracy of the dual-input DBN on AURORA-2 as a function of state-likelihood-dimensionality of SC input, with a trained CPT. The left hand panel (a) shows the data for test set A, the right hand panel for test set B. Note that the horizontal axis is displayed logarithmically. The data point on the far right represents 179 likelihood dimensions.

and state probability estimates. Another difference is that the optimal number of SC dimensions for test set A is now smaller than for test set B: Since on test set A, the probability estimates are more likely to be correct in noisy speech due to matched noise dictionary, the CPT that was trained on clean speech can also be helpful in noisier speech, leading to higher accuracies at lower N's. For test set B, the probability estimates on noisy speech resemble the clean speech probability estimates less, and thus the use of a trained CPT is less effective.

6. CONCLUSIONS

In this work, we proposed two methods to further improve the noise robustness of the ASR system described in [6], which consists of a dual-input DBN simultaneously observing MFCC acoustic features modelled by a GMM and an exemplar-based Sparse Classification (SC) state probability stream. The first modification was the integration of more contextual information by increasing the exemplar size from 10 to 30 frames. The second modification was increasing the dimensionality of the SC input stream from the first-best state probability to N-best probability estimates.

Experiments on AURORA-2 revealed that incorporating more time context improved the noise robustness substantially, especially at low SNRs. At the same time, unlike when using only the SC input stream, the dual-input DBN did not suffer a loss in performance when using more time context.

In a DBN only observing the SC input stream it was found that, when the dimensionality of the vector with the state-probability estimates is increased, the recognition performance typically increases, although the accuracy quickly converges to a constant value as additional state probability estimates do not carry any additional information. In the dual-input DBN system, increasing the dimensionality of the state-probability estimations also improved the recognition accuracy on noisy speech, although there was a clear optimum: the use of too many SC state-probability estimates opens up the search space too much, allowing erroneous GMM-based probability estimates to have a detrimental effect.

In general, it was found the dual-input DBN offered a greatly improved noise robustness, while retaining the high performance on clean speech offered by the GMM-based modelling of acoustic features. In future work, we will investigate the effect of using trained CPT's for the SC input stream, which will allow the system more flexibility to handle the provided SC state probabilities, as well as the use of multi-condition training data.

REFERENCES

[1] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The development of the AMI system for the transcription of speech in meetings," in *2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.

[2] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.*, vol. 22, no. 1, pp. 1–15, 1997.

[3] P .Divenyi, *Speech separation by humans and machines*, Kluwer, 2005.

[4] H. Bourlard and N. Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *International School on Neural Nets: Adaptive Processing of Temporal Information*. 1997, Springer Verlag.

[5] V. Goel, S. Kumar, and W. Byrne, "Segmental minimum bayes-risk decoding for automatic speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 12:234 – 249, 2004.

[6] Y. Sun, J.F. Gemmeke, B. Cranen, L. ten Bosch, and L. Boves, "Using a DBN to integrate sparse classification and gmm-based ASR," in *Proceedings of Interspeech 2010*, Makuhari, Japan, 2010.

[7] J.F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proceedings of ICASSP 2010*, Dallas, USA, 2010.

[8] J. Bilmes, "On soft evidence in bayesian networks," Tech. Rep. UWEETR-2004-0016, University of Washington, Dept. of Electrical Engineering, 2004.